Could DNA be the next big thing in data storage?

DNA storage technology is advancing rapidly. Atos is closely monitoring the development of this technology in the hope that future developments will make it a scalable, cost effective and crucially a decarbonized alternative to today's carbon intensive technologies.

Trusted partner for your **Digital Journey**

Thought Leadership

Atos

Could DNA be the next big thing in data storage?

DNA storage sounds a bit like science fiction, but the quest to make DNA a storage medium has been around much longer than you might think. This paper details the current state of play in DNA storage and considerations organizations should make for it.

What might DNA storage look like?

DNA digital data storage ("DNA storage" in short) is the process of encoding and decoding digital binary data to and from synthesized or native strands of DNA the building block of life as we know it. DNA is made from four types of components (called bases): adenine (A), thymine (T), guanine (G) and cytosine (C). DNA is a code made of sequences of As, Ts, Gs and Cs instead of 1s and 0s. Furthermore, DNA is very densely packed: 3 billion base pairs fit into 6 microns, a space about one tenth of the thickness of a human hair.

Types of technologies

There are cell-free and cell-based (in vivo) technologies.

Cell-free technologies

- Dye sequencing: DNA is sequenced by immobilizing single strands, amplifying the sequence with polymerase chain reactions (PCR), labelling the bases and tagging the complementary bases. The resulting fluorescence pattern for bases is captured as an image and processed to determine the DNA sequence.
- Nanopore: DNA molecules pass through nanopores, causing measurable changes in electrical resistivity. This enables detection in real time, but the accuracy is not yet good enough for data storage purposes.

In vivo technologies:

The genetic code of a living organism is co-opted for data storage and retrieval. This approach can be extended with synthetic means such as engineering cells with "molecular recorders" or using CRISPR to insert artificial DNA sequences into the cell's genome.

Why DNA storage

Estimates for the total amount of accumulated data in the world vary between 40 and 60 ZB, that is, trillions of GBs. Today, much of that is stored in 8.2 million data centers the largest of which is 7.2 million sq feet, about the equivalent of 147 soccer pitches. The rate at which humanity and machines are producing data continues to increase exponentially, so the figure increases dramatically year after year.

But there is also a preservation issue. We typically store data in media that cannot last more than a few decades, and therefore needs frequent checks and replacement (cost issue) and which reside or are processed in billions of devices and huge data centers throughout the Cloud continuum (cost and sustainability issue).

As a result, there is a growing gap between our capacity to generate information and our ability to store it adequately – for long periods of time, with fast write and read times, in a cost-efficient manner and in a sustainable way, even in extreme conditions.

DNA is a very high-density medium of information storage, in a very durable way (even under adverse environmental conditions) and does not require much energy (it can be stored at low temperatures but also in a vitreous state – dried). Therefore, it has been proposed as the digital data storage medium of the (near?) future.

History of DNA storage: from 1959 to Present day

The concept of DNA storage originates in the late 1950s – shortly after the publication of Watson and Crick's double-helix model of DNA – and during the 1960s, from luminaries like Feynmann, Neiman and Wiener.

In the 1980s there was a scientific-artistic experiment in which the image of a very simple character (a rune), encoded as a 5x7 matrix, was stored in a DNA sequence of the bacteria E. coli.

The 2010s saw an explosion of results:

- In 2011, Church and colleagues at Harvard encoded a 659 KB book, which they repeated with a larger book in 2012.
- In 2013, the European Bioinformatics Institute stored, retrieved and reproduced more than 5 million bits with 99,99% accuracy. They introduced the use of error-correcting encoding schemes (usual in information and communication technology) and the use of overlaps and redundancy.
- In 2015, ETH Zurich reported the long-term stability of data encoded in DNA.
- In 2017, scientists at Columbia University and the New York Genome Center published the method known as DNA Fountain, which enables the maximum theoretical information channel capacity to be approached, albeit at a high cost.
- In 2018, researchers at the University of Washington and Microsoft Research developed a fully automated system for writing, storing and reading data encoded in DNA.
- In January 2019, researchers showed how to encode relational data in synthetic DNA and perform – chemically – data processing operations similar to SQL.
- In June 2019, scientists reported that all 16 GB of text from Wikipedia's English-language version have been encoded into synthetic DNA.
- In 2020, an article described DNA storage in native DNA sequences, as well as a method enabling bit-wise random access and in-memory computing features.

DNA computing is slow per computation but capable of highly parallelized computations. Therefore, if DNA computing matures beyond its current emerging state, it might rival quantum computing in some respects and have a positive impact on decarbonization.

Today, DNA can be read (sequenced), written (synthesized) and accurately copied. Shannon's limit (the rate at which data can be transferred in an error free manner) for channel capacity (2 bits per nucleotide) is nearly reached (results range from 85% to 98%). However, at present, DNA Storage is not a realistic alternative due to its low speed and high cost.

Researchers are currently focusing on aspects such as optimal error correction schemes or optimal mapping strategies – to convert digital data into DNA sequences within the biochemical constraints. Error correction schemes are vital, since the natural proof-reading and repair mechanisms that enzymes provide are not available in synthesized DNA.

Some researchers are attempting to go beyond Shannon's limit by using degenerate bases (instead of the four basic ones: A, T, C, G) but, at least for the moment, the approach has not yielded positive results.

In addition to the work aiming at discovering new features and possibilities, there are efforts directed at making existing features ready for large-scale production. Some of the features we've seen in the previous section are very promising but are either very slow, not easily repeatable or very costly.

What is the potential promise of this technology?

- Very high density
- Low energy intensity
- Low carbon usage
- High durability and preservation.
- Innovations in biotechnology move rapidly:
  - Thanks to advances in genome sequencing, it is already possible to read billions of DNA sequences in parallel.
  - Scientists already use “bar coding” (using DNA sequences as molecular ID tags) to keep track of experimental results.
  - Advances of mRNA technologies for COVID-19 have focused attention on biotechnology science.

There are, however, important limitations that impede its widespread use for the moment:

- Write and read times are still quite slow compared to optical or electromagnetic media.
- The cost is still very high, between 10^7 and 10^8 times that of a modern 1 TB hard drive. It took about 40 years for conventional, large-capacity disks to achieve a price reduction of a factor of 10^7, although increasingly fast technological developments may reduce that time gap significantly for DNA storage.

Key take-aways

DNA storage is a very promising technology, as a very dense, durable and efficient storage medium. Very significant advances have taken place during the past decade. If the cost and complexity of the technology decrease during the 2020s, it may become a viable alternative to current technologies for business purposes. The achievements in cost, speed and carbon impact will determine the specific business use cases where the technology can be applied.

DNA storage will likely enable very high densities of information with – hopefully – increasingly fast read and write speeds, the technology will pull from the other two dimensions of the “AI triumvirate”: computing and algorithms. We could expect significant milestones in algorithms that enable complex processing of very large amounts of data, as in bioengineering but for other domains (retail, energy etc.), as well as increased demand for HPC and quantum computing.

As a speculative concept, some researchers have introduced the “DNA of Things” (DoT): digital data can be encoded in DNA molecules that would be embedded in objects. The objects would then carry their own blueprint, like living organisms. Where IoT is about interrelated computing objects, DoT would be about independent storage objects (although dependencies could be devised via replication and other data exchanges).

If / when the technology becomes available, it will mark a “before and after” moment in digital technology, shaping the next information space.

Atos Clients are advised to add DNA data storage to their long term outlook and monitor Atos output related to the topic.

What will Atos do?

Atos is actively watching developments and research in the field of DNA storage - with particular focus on our strategic objectives of decarbonized digital. While it’s too early at this point for inclusion in our portfolio we’ve added and incorporated DNA storage into our vision for data storage technologies and look forward to playing an active role in experimentation and innovation when the technology matures more. We’ll keep our clients abreast of the developments in this field with ad-hoc updates from our Atos Scientific Community.

Jose Esteban Lauzan
Author Atos Scientific Community

John Hall
Editor in Chief Atos Scientific Community

Mark Smith
Editor in Chief Atos

Sources

- https://en.wikipedia.org/wiki/DNA_digital_data_storage
- https://www.scientificamerican.com/article/dna-data-storage-is-closer-than-you-think/
- https://www.nature.com/articles/s41576-019-0125-3
- https://link.springer.com/article/10.1007/s11033-018-4280-y
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5935598/
- https://www.cnet.com/news/startup-packs-all-16gb-wikipedia-onto-dna-strands-demonstrate-new-storage-tech/
- https://jbioleng.biomedcentral.com/articles/10.1186/s13036-019-0211-2

About Atos

Atos is a global leader in digital transformation with 105,000 employees and annual revenue of over € 11 billion. European number one in cybersecurity, cloud and high performance computing, the Group provides tailored end-to-end solutions for all industries in 71 countries. A pioneer in decarbonization services and products, Atos is committed to a secure and decarbonized digital for its clients. Atos operates under the brands Atos and Atos|Syntel. Atos is a SE (Societas Europaea), listed on the CAC40 Paris stock index.

The purpose of Atos is to help design the future of the information space. Its expertise and services support the development of knowledge, education and research in a multicultural approach and contribute to the development of scientific and technological excellence. Across the world, the Group enables its customers and employees, and members of societies at large to live, work and develop sustainably, in a safe and secure information space.

The Atos Scientific Community crafts the Group's vision for the future of technology in business, and anticipates the upcoming trends and technologies that will shape businesses and society in the years ahead.

Find out more about us
atos.net
atos.net/scientific-community

Let's start a discussion together

For more information: **atos.net/en/contact-us**

CT-210706-SKR-WP-TechDays-DNA